Data Analysis and Visualization Exercise 10

Matthias Heinig, Felix Brechtmann, Daniela Klaproth-Andrade, Julien Gagneur

Section 00 - Getting Ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)
library(tidyr)
library(dplyr)
library(patchwork) # optional, makes plots nicer
```

Section 01 - Linear regression for Predicting Heights

To start, read the provided heights dataset using the following line of code (it's your own heights data):

```
heights <- fread("extdata/height.csv") %>% na.omit() %>%
  .[, sex:=as.factor(toupper(sex))]
heights
```

1. Predict each student's hight, given their sex and their parents heights.

2. Check the plot of the residual vs the predicted values and the Q-Q plot of the residuals. Do these plots provide evidence against the assumptions of linear regression?

3. Let's focus on one sex. Fit a linear model for each male's height given the father's height. Then, fit another linear model for the father's height given each male's height. To do so, subset the table to only include male students.

4. Predict each student's height given the father's height (`predict()`) and predict each father's height given each student's height. Store both predictions into new columns of a data table. Then, plot the original data and additionally both regression lines.

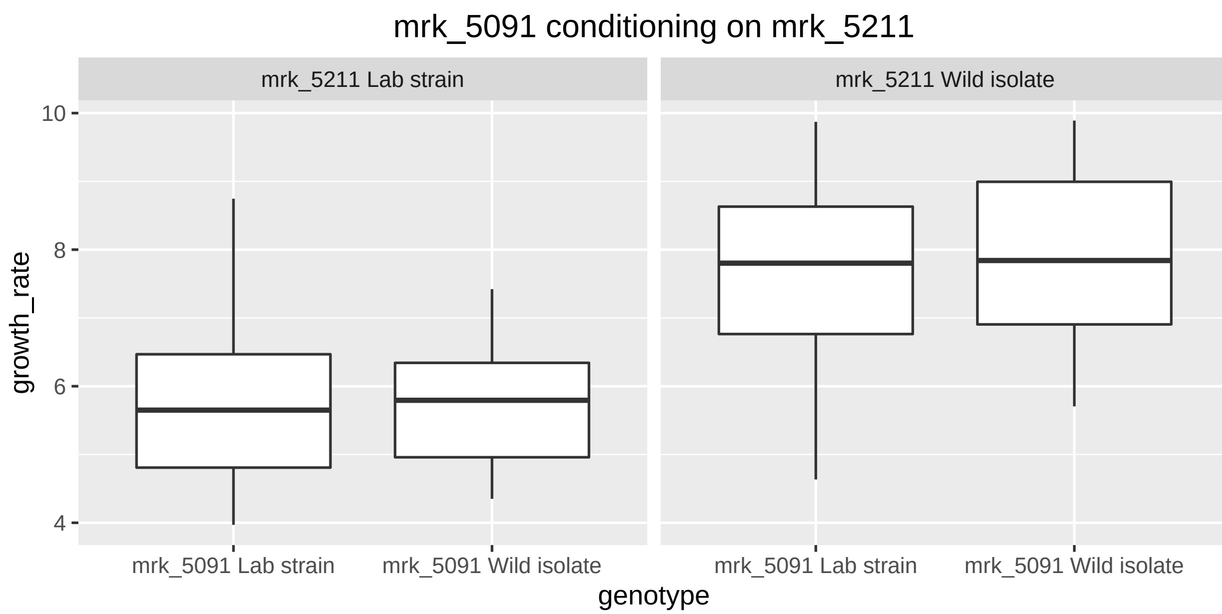
5. Additionally, run a PCA on the same subset of the data and plot the first principal component into the same figure.

6. Interpret the plot from above. How can we explain the different slopes of the two linear models and the pca?

Section 02 - Adjusting for confounding variables

Yeast QTL

Recall the yeast QTL data set from the previous exercises and lecture. In particular, we consider once again the question: Does marker 5091 still associate with growth in maltose when conditioned on marker 5211? Here is the plot of the data:




```r
growth <- fread(file.path(eqtl_dir, "growth.txt"))
growth <- growth %>% melt(id.vars="strain", variable.name='media', value.name='growth_rate')
growth <- growth[media=="YPMalt"]

genotype <- fread(file.path(eqtl_dir, "genotype.txt"))
genotype <- genotype[, .(strain, mrk_5211, mrk_5091)]

head(genotype)
```

```
##     strain     mrk_5211     mrk_5091
## 1: seg_01B   Lab strain   Lab strain
## 2: seg_01C   Lab strain   Lab strain
## 3: seg_01D Wild isolate Wild isolate
## 4: seg_02B   Lab strain Wild isolate
## 5: seg_02C   Lab strain   Lab strain
## 6: seg_02D Wild isolate   Lab strain
```

```r
head(growth)
```

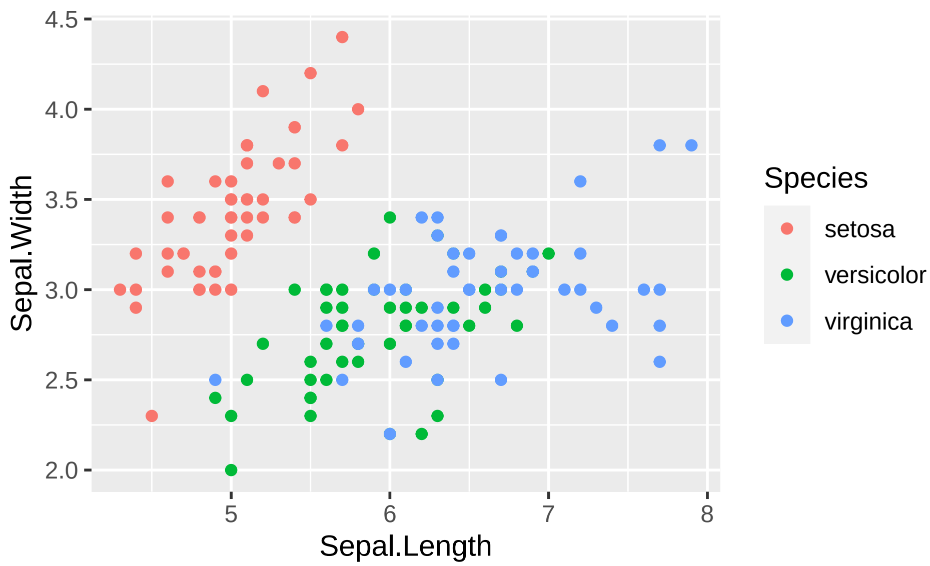
```
##     strain  media growth_rate
## 1: seg_01B YPMalt    6.720447
## 2: seg_01C YPMalt    7.429273
## 3: seg_01D YPMalt    6.905589
## 4: seg_02B YPMalt    4.924324
## 5: seg_02C YPMalt    4.413402
## 6: seg_02D YPMalt    7.926200
```

1. Run a linear model predicting the growth given both genotypes and interpret the result. Call this model `full`.

2. Create a reduced model that only depends on the genotype of `mrk_5211`. Then run anova to compare the full and the reduced model. What do you conclude?

Iris dataset

Recall the plot showing the relationship between sepal width and sepal length in the Iris dataset:



1. Fit three linear models predicting Sepal.Width from Sepal.Length: the base model that simply predicts sepal width from sepal length, one where you use the species as a covariate in linear regression (i.e., different intercept for different species) and one where you use separate slopes and intercepts for different species (Hint: use the `*` operator in `lm`: `lm(y ~ x1 * x2)`)

2. Overlay the resulting fits on the plot above (Hint: use `predict` to generate predictions)

3. Use `anova` to test if the second model is a better model than the base and also if the third model is better than the second.

Section 03 - Simulation of the estimates of β

Assume a simple linear model with parameters $\alpha = -0.5, \beta = 1.5, \sigma^2 = 0.7$.

- First simulate a fixed set of $N = 500$ values for x from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 20$.
- Then, run the simulation 1000 times. Each time, draw the values of y according to the model specification $y = \alpha + \beta x$, estimate a linear model from the data and store the value of $\hat{\beta}$.
- Finally, create a histogram of the values of $\hat{\beta}$. Indicate the true β and the mean of $\hat{\beta}$ by vertical lines. Also indicate the theoretical distribution of β as a curve.

[Optional] Section 04 - Perform a simulation for nested models

Assume two nested linear models:

- one reduced model with parameters $\beta_0 = -0.5, \beta_1 = 1.5, \sigma^2 = 0.7$
- full model with parameters $\beta_0 = -0.5, \beta_1 = 1.5, \beta_2 = -0.5, \sigma^2 = 0.7$

1. Simulate a fixed set of $N = 500$ values for x_1 and for x_2 from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 20$. Run the simulation 1000 times. In each simulation: * draw the values of y according to the reduced model * fit the reduced and the full model to the simulated data * compute the F statistic of the model comparison

2. Create a histogram of the simulated F statistics. Plot the theoretical distribution of the F statistic on top of the histogram

3. Draw values of y according to the full model (once). Fit the reduced and the full model to the simulated data. Compute the F statistic for the model comparison. Compare the F statistic from the full model to the distribution of F statistics from the reduced model